EDA-IPL

Submitted by:

Akanksha Jha[2024010007]

Aryan Kamboj[2024010018]

MCA 2nd Year

Submitted to:

Dr. Anjula Mehto

Assistant Professor

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Computer Science and Engineering Department

Thapar Institute of Engineering and Technology, Patiala

November 2025

TABLE OF CONTENTS

Introduction or Project Overview

The Indian Premier League (IPL) is one of the most dynamic and data-rich cricket tournaments in the world, featuring fast-paced matches, diverse teams, and continuously evolving strategies. With thousands of deliveries bowled and hundreds of matches played each season, IPL data serves as a valuable resource for analyzing team performance, player behavior, and match outcomes. To utilize this potential, the present project focuses on performing an extensive Exploratory Data Analysis (EDA) on historical IPL match and ball-by-ball datasets.Through this project ,the goal is to extract meaningful insights from datasets by pattern identification, and graphical visualizations.

By performing this EDA, the project provides a comprehensive overview of IPL match dynamics over more than a decade. It not only highlights statistical patterns but also strengthens understanding of how small events at the ball level accumulate to shape outcomes at the match level. Ultimately, this analysis establishes a strong foundation for further advanced studies such as predictive modeling, player performance forecasting, or strategy optimization in cricket analytics.

Problem Statement

Although the IPL datasets contain extensive match-level and ball-by-ball information, the raw data does not directly reveal underlying trends or influencing factors within the tournament. The challenge is to identify meaningful patterns hidden within large volumes of numerical and categorical data and transform them into actionable cricket insights.

This project addresses the need to systematically explore IPL data to answer key analytical questions related to team performance, player contributions, and match outcomes. Specifically, it aims to determine:

- Which cities and stadiums host the most matches, and how venue distribution influences the tournament.
- Which players consistently perform at award-winning levels, based on Player of the Match statistics.
- How toss decisions affect winning probability, and whether choosing to bat or field first offers an advantage.
- Which types of extras occur most frequently—wides, no-balls, byes, and leg-byes—and which bowlers concede them most often.
- How players are dismissed across seasons, including which batsmen are dismissed most by specific bowlers or fielders, and the dominant dismissal types.
-

Overall, the problem is to convert raw IPL match and ball-by-ball records into clear, interpretable insights that explain how various factors influence match outcomes and player behavior. Achieving this requires careful data cleaning, aggregation, visualization, and trend comparison across multiple seasons

Overview of the Dataset used

The dataset used in this project consists of two primary CSV files spanning 13 IPL seasons from 2008 to 2020. The first file, *"IPL Matches 2008–2020.csv"*, contains 816 match records, including information on match venues, participating teams, toss outcomes, winners, umpires, and result margins.

city: City where the match was played.
date: Match date.
player_of_match: Player awarded Man of the Match.
venue: Stadium where the match took place.
neutral_venue: Neutral venue indicator (0/1).
team1: First competing team.
team2: Second competing team.
toss_winner: Team that won the toss.
toss_decision: Toss decision (bat/field).
winner: Team that won the match.
result: Match result type (normal, tie, no result).
result_margin: Margin of victory.
eliminator: Indicates if the match was an eliminator.
method: D/L method applied (if applicable).
umpire1: First on-field umpire.
umpire2: Second on-field umpire

The second file, *"IPL Ball-by-Ball 2008–2020.csv"*, provides a much deeper level of granularity, with 193,468 ball-level entries capturing every delivery bowled across all seasons. This includes details such as the batsman and bowler names, runs scored, extras conceded, wickets taken, dismissal types, and fielding involvement

inning: Inning number (1/2).
over: Over number.
ball: Ball within the over.
batsman: Striker facing the delivery.
non_striker: Batsman at the other end.
bowler: Bowler delivering the ball.
batsman_runs: Runs scored off the bat.
extra_runs: Runs given as extras.
total_runs: Total runs for the delivery.
non_boundary: Indicates non-boundary shot (0/1).
is_wicket: Whether a wicket fell (0/1).
dismissal_kind: Type of dismissal, if any.
player_dismissed: Dismissed player's name.
fielder: Fielder involved in dismissal.
extras_type: Type of extra awarded.
batting_team: Team batting.
bowling_team: Team bowling.

Project Workflow

The major steps performed are described in detail below.

1. Loading the Datasets
The analysis began with importing the required Python libraries, including Pandas, NumPy, Matplotlib, and Seaborn. Two datasets were then loaded into Pandas DataFrames and the initial few rows of both datasets were inspected to confirm successful loading and to understand the structure of the data.

2. Data Cleaning and Preliminary Inspection
Basic data cleaning steps were performed, starting with the removal of the unnecessary id column from both datasets. A thorough structural inspection using .info() revealed:

- Several columns contained missing values, especially in attributes such as city, player of match, winner, result margin, and particularly method, which had only a small number of non-null entries.
- In the ball-by-ball dataset, dismissal-related columns (such as dismissal kind, player dismissed, and fielder) contained many null values, since most deliveries do not involve a wicket.

The unique values of key categorical fields such as city, method, and extras_type were examined to understand data variability and potential inconsistencies. Neutral venue matches were also identified, revealing 77 matches played at neutral locations during certain IPL seasons.

3. Descriptive Statistics and Frequency Counts
The next phase involved examining high-level patterns through descriptive statistics and frequency distributions:

- Player of the Match counts were computed to identify the top performers.
- City-wise and venue-wise distributions were analyzed to understand which locations hosted the largest number of matches.
- Toss decisions and match outcomes were explored using grouped views and frequency tables.
- In the ball-by-ball dataset, summary statistics were generated to understand the numerical distribution of overs, balls, batsman runs, extra runs, and total runs.
- Count distributions were produced for extras types, including wides, no-balls, byes, leg-byes, and penalties.

4. Visual Exploratory Analysis

A substantial portion of the workflow involved creating visualizations to uncover patterns more effectively:

- Bar charts were generated to show the top players of the match, top cities, and top stadiums.
- Countplots were created to compare match winners based on toss decisions.
- A distribution plot of result margins illustrated how closely or comfortably matches were won.
- In the ball-by-ball dataset, detailed barplots and pie charts were produced to examine:
  - which bowlers conceded the highest numbers of wides, no-balls, and leg-byes,
  - which batsmen were associated with the most leg-byes,
  - the proportions of different extras types across all seasons.

5. Focused Queries and Deep-Dive Analysis

Further analysis was performed using specific queries and grouped operations to answer deeper cricket-related questions:

- A breakdown of caught dismissals was created, identifying batsman–fielder combinations and the frequency of each.
- Individual player analyses were conducted, such as identifying bowlers involved in the run-out dismissals of high-profile batsmen like AB de Villiers and Virat Kohli..
- Comparisons were made between toss winners and match winners, especially for matches where the toss decision was to field first, enabling insights into the tactical impact of the toss.

6. Extras-Type and Bowler Behavior Analysis

A specialized section of the workflow focused on extras, as these runs often have a significant impact on match outcomes. The project:

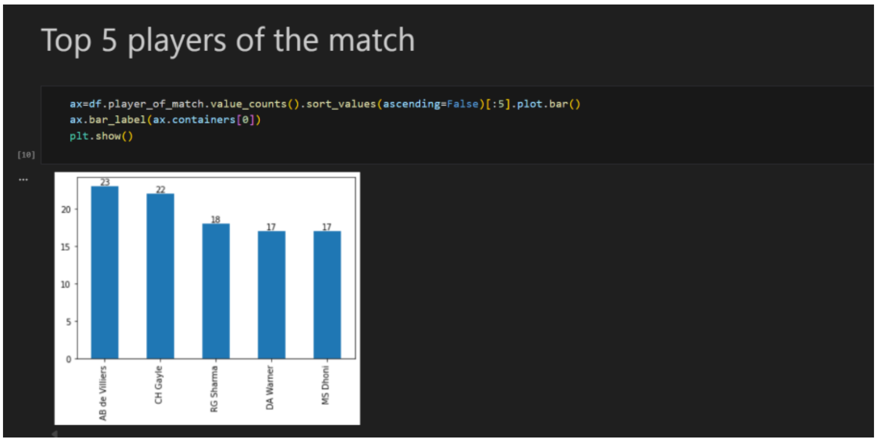
- Categorized extras into byes, leg-byes, wides, no-balls, and penalties.
- Identified the top bowlers conceding each category of extras.
- Examined extras conceded by individual bowlers such as Amit Mishra, Zaheer Khan, and YS Chahal, highlighting patterns in their bowling discipline.

7. Dismissal Pattern Investigation

The workflow also examined wicket-taking patterns by analyzing:

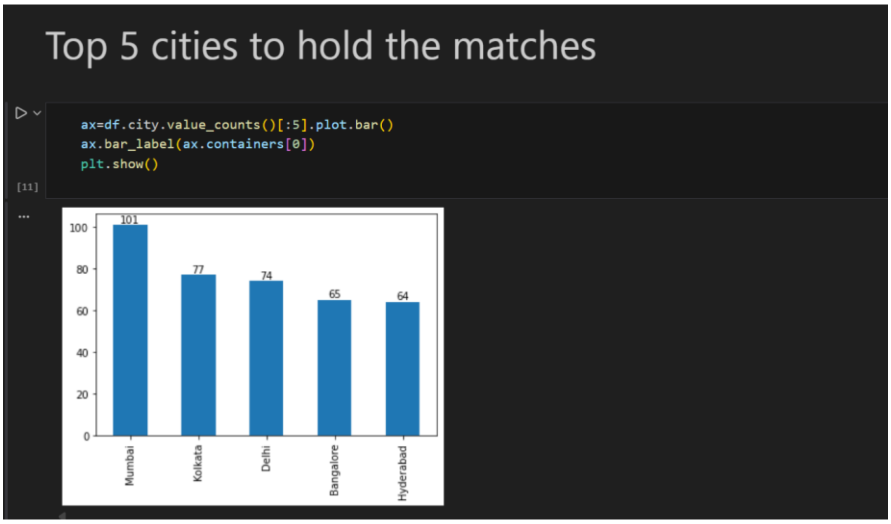
- The various types of dismissals recorded (caught, run out, bowled, lbw, stumped, etc.).
- The frequency and combinations of caught dismissals, including the most recurring fielder–batsman pairs.
- Specific investigations into run-out scenarios for star players, revealing which bowlers or fielders frequently contributed to their dismissals.

Results



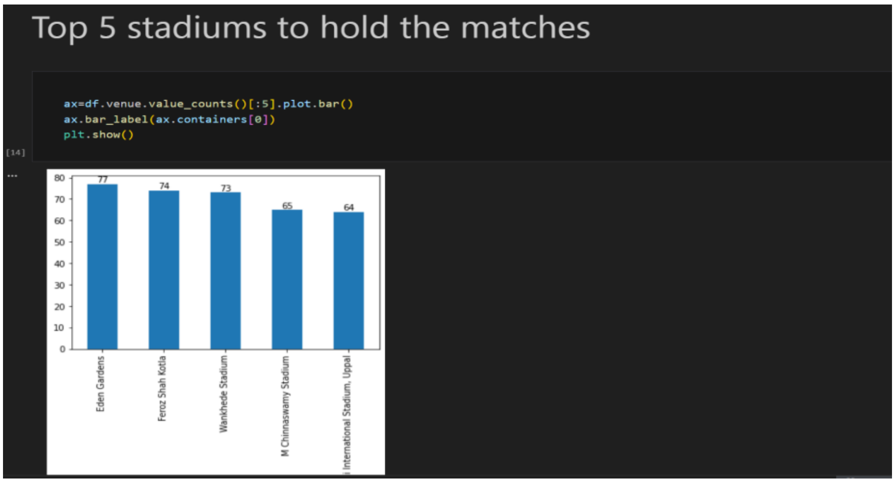
Top 5 players of the match

```
ax=df.player_of_match.value_counts().sort_values(ascending=False)[:5].plot.bar()
ax.bar_label(ax.containers[0])
plt.show()
```



Top 5 cities to hold the matches

```
ax=df.city.value_counts()[:5].plot.bar()
ax.bar_label(ax.containers[0])
plt.show()
```

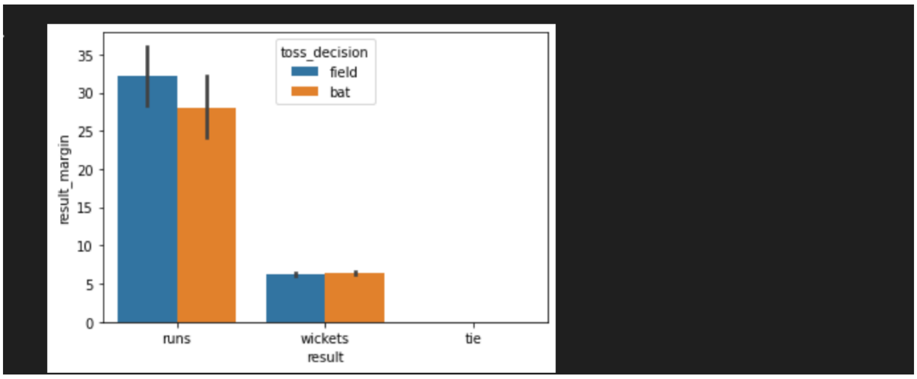


Top 5 stadiums to hold the matches

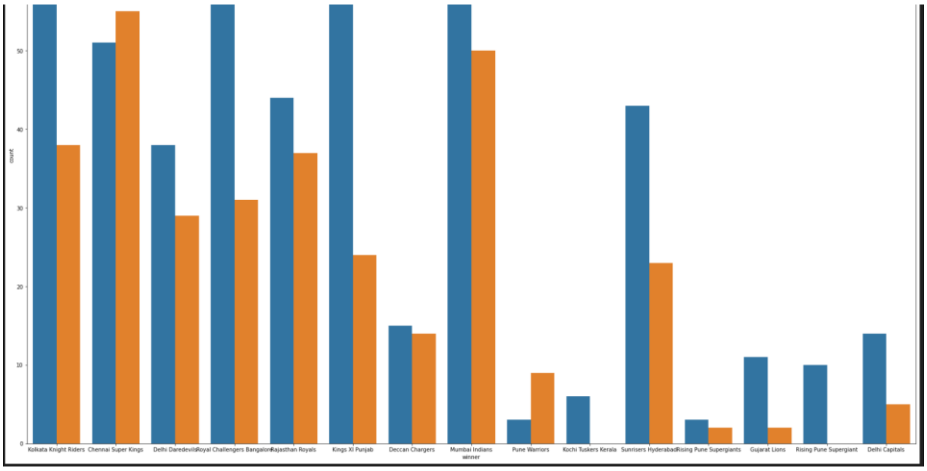
```
ax=df.venue.value_counts()[:5].plot.bar()
ax.bar_label(ax.containers[0])
plt.show()
```

[14]

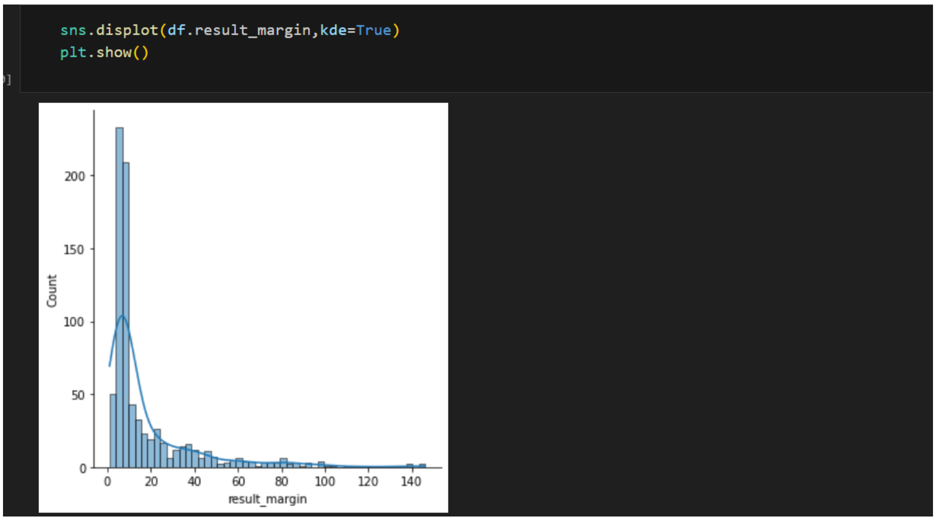
Toss decision vs result margin



countplot of winners split by toss decision

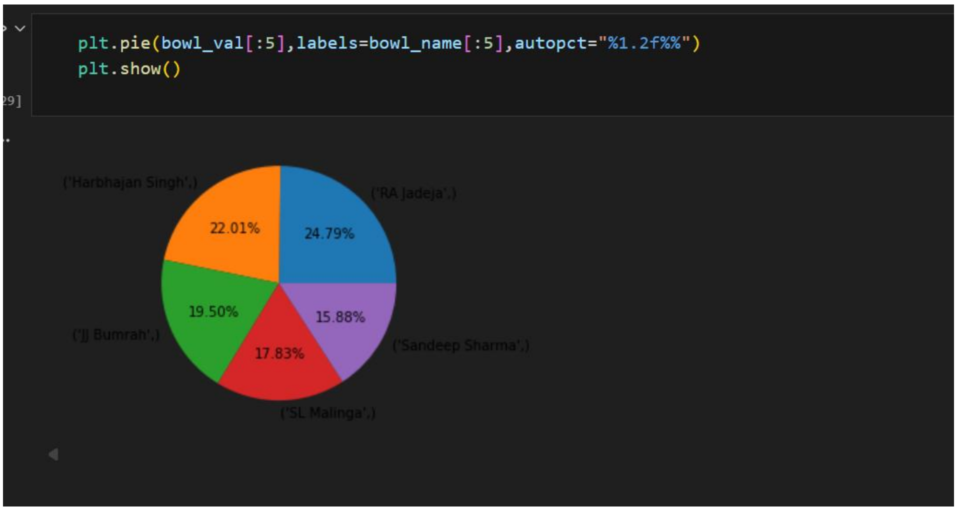


distribution plot for result_margin



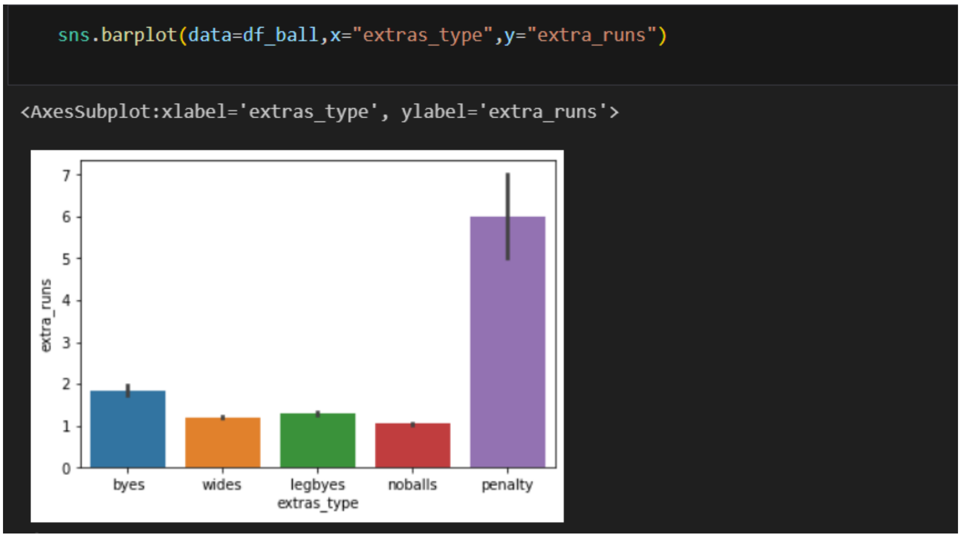
```
sns.displot(df.result_margin,kde=True)
plt.show()
```

top-5 bowlers against AB de Villiers



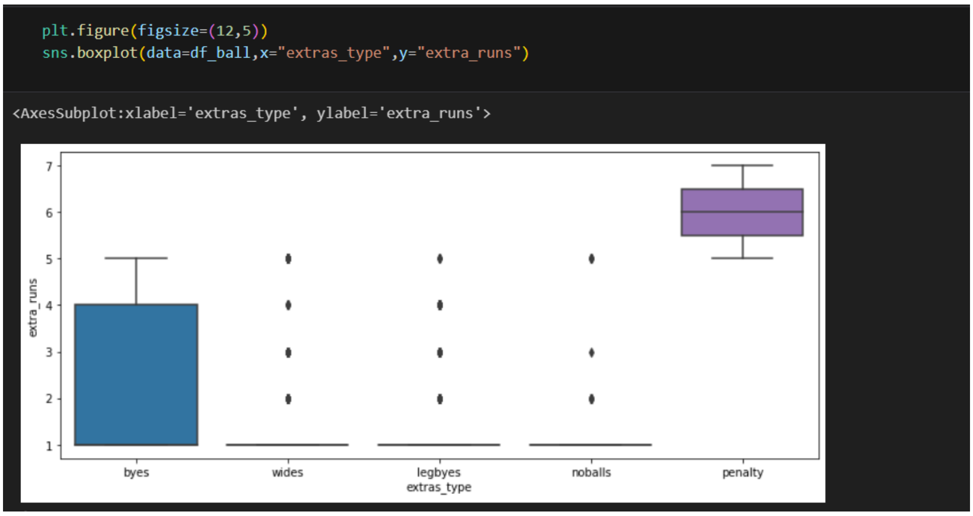
```
plt.pie(bowl_val[:5],labels=bowl_name[:5],autopct="%1.2f%%")
plt.show()
```

extras types vs extra_runs



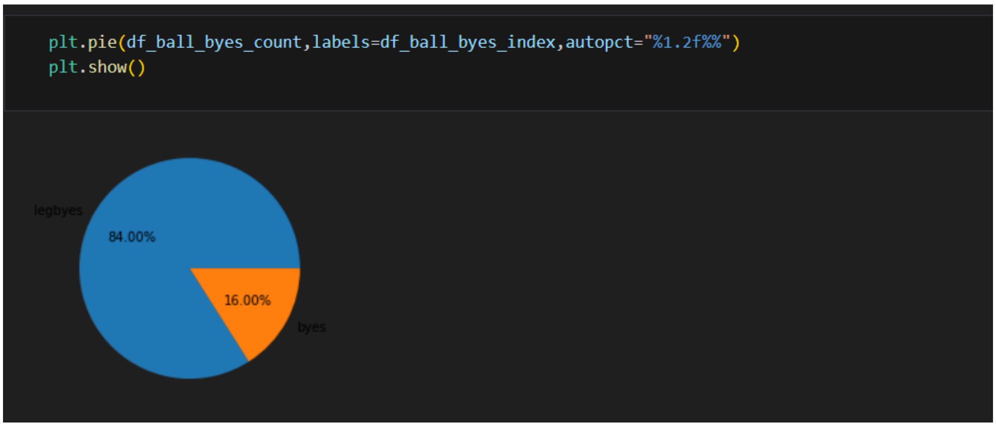
```
sns.barplot(data=df_ball,x="extras_type",y="extra_runs")
```

```
<AxesSubplot:xlabel='extras_type', ylabel='extra_runs'>
```

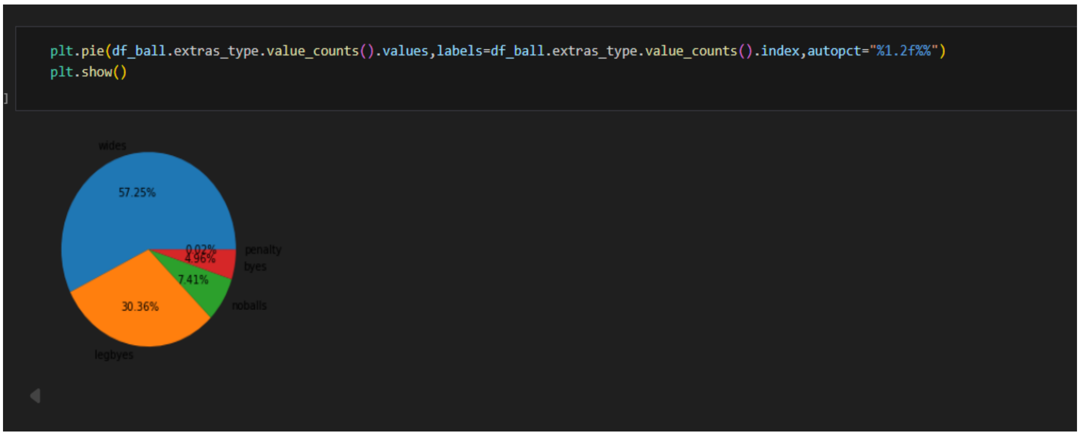


```
plt.figure(figsize=(12,5))
sns.boxplot(data=df_ball,x="extras_type",y="extra_runs")
```

```
<AxesSubplot:xlabel='extras_type', ylabel='extra_runs'>
```

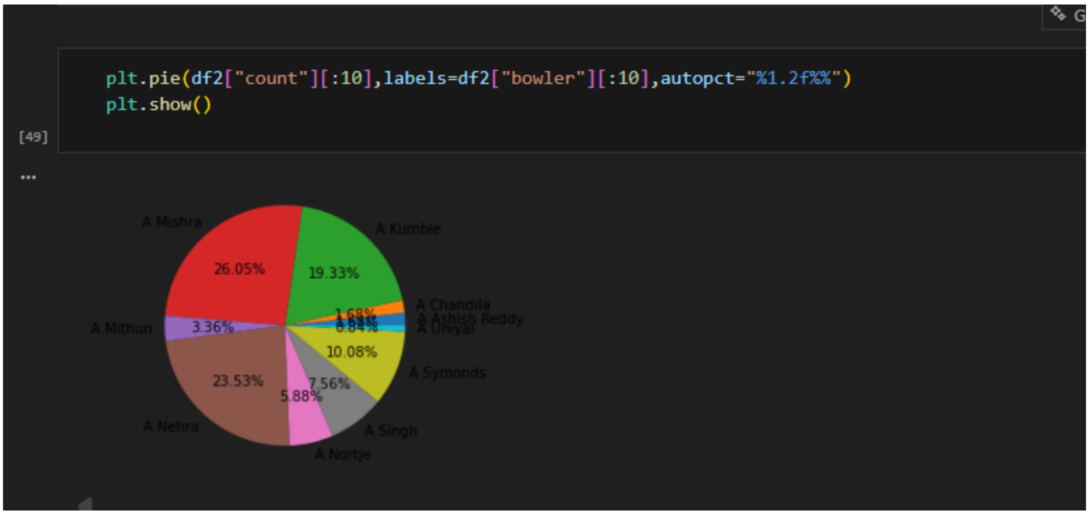


```
plt.pie(df_ball_byes_count,labels=df_ball_byes_index,autopct="%1.2f%%")
plt.show()
```

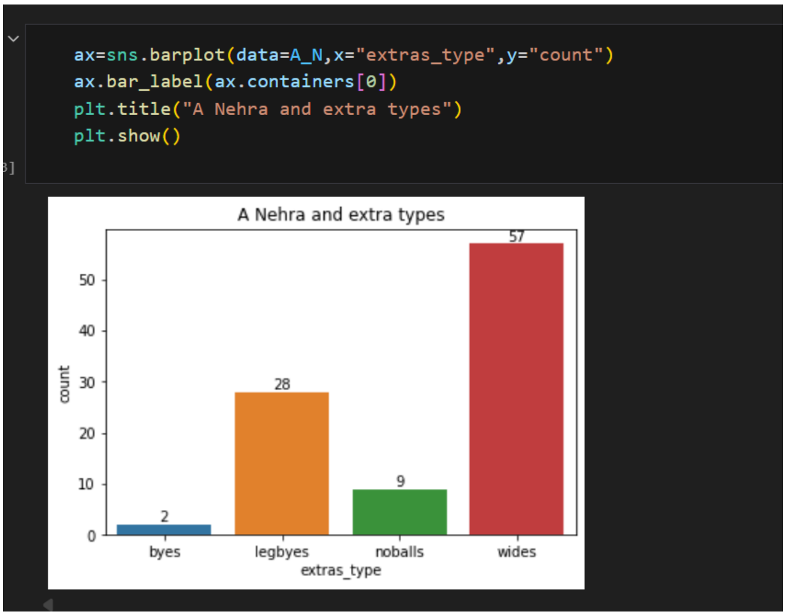


```
plt.pie(df_ball.extras_type.value_counts().values,labels=df_ball.extras_type.value_counts().index,autopct="%1.2f%%")
plt.show()
```

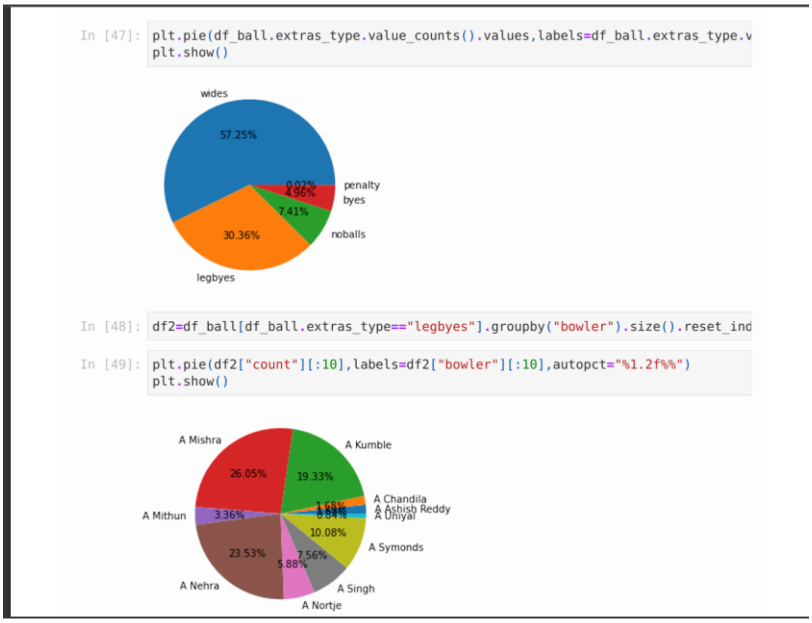
Group legbyes by bowler



```
plt.pie(df2["count"][:10],labels=df2["bowler"][:10],autopct="%1.2f%%")
plt.show()
```



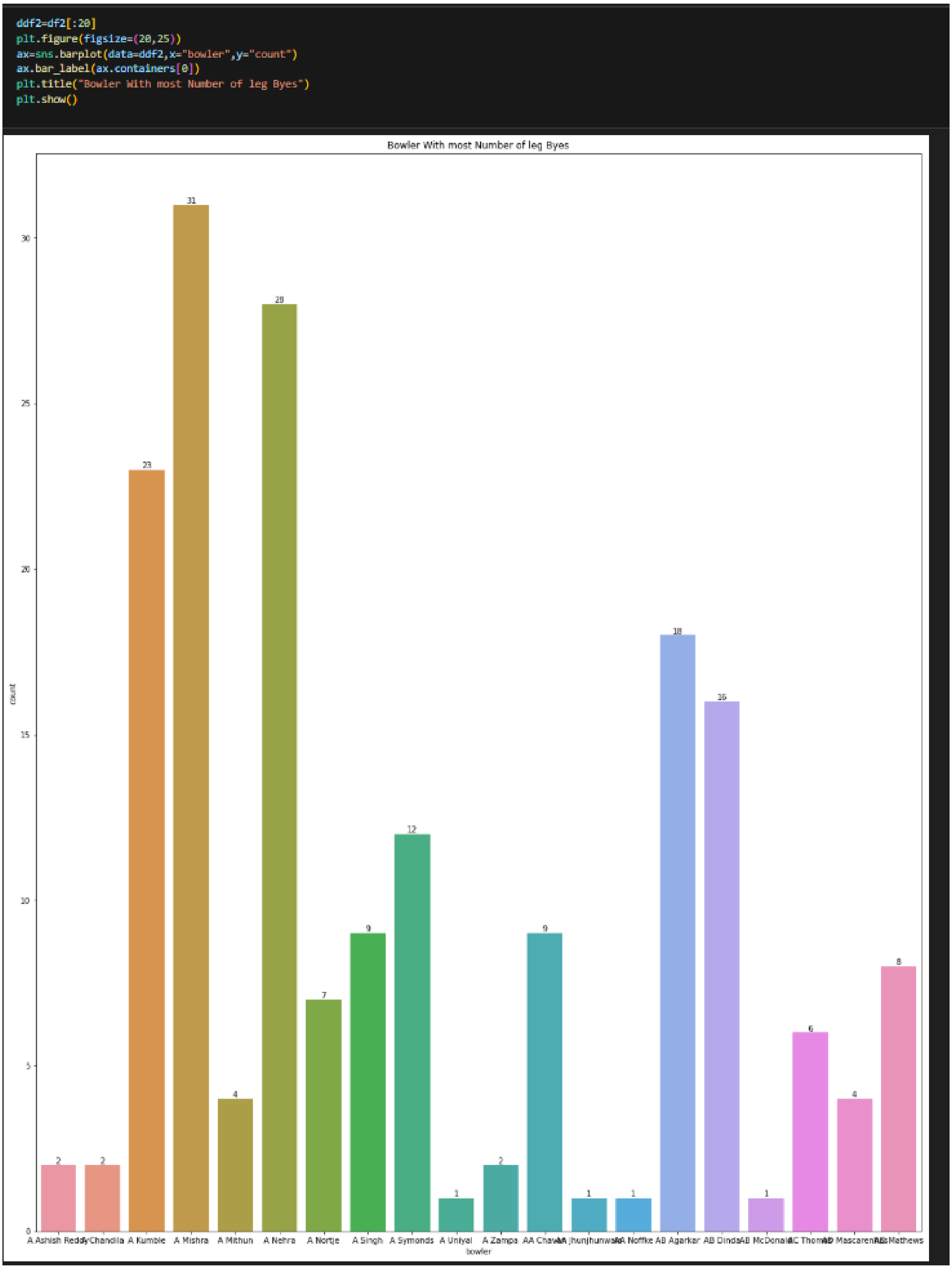
```
ax=sns.barplot(data=A_N,x="extras_type",y="count")
ax.bar_label(ax.containers[0])
plt.title("A Nehra and extra types")
plt.show()
```



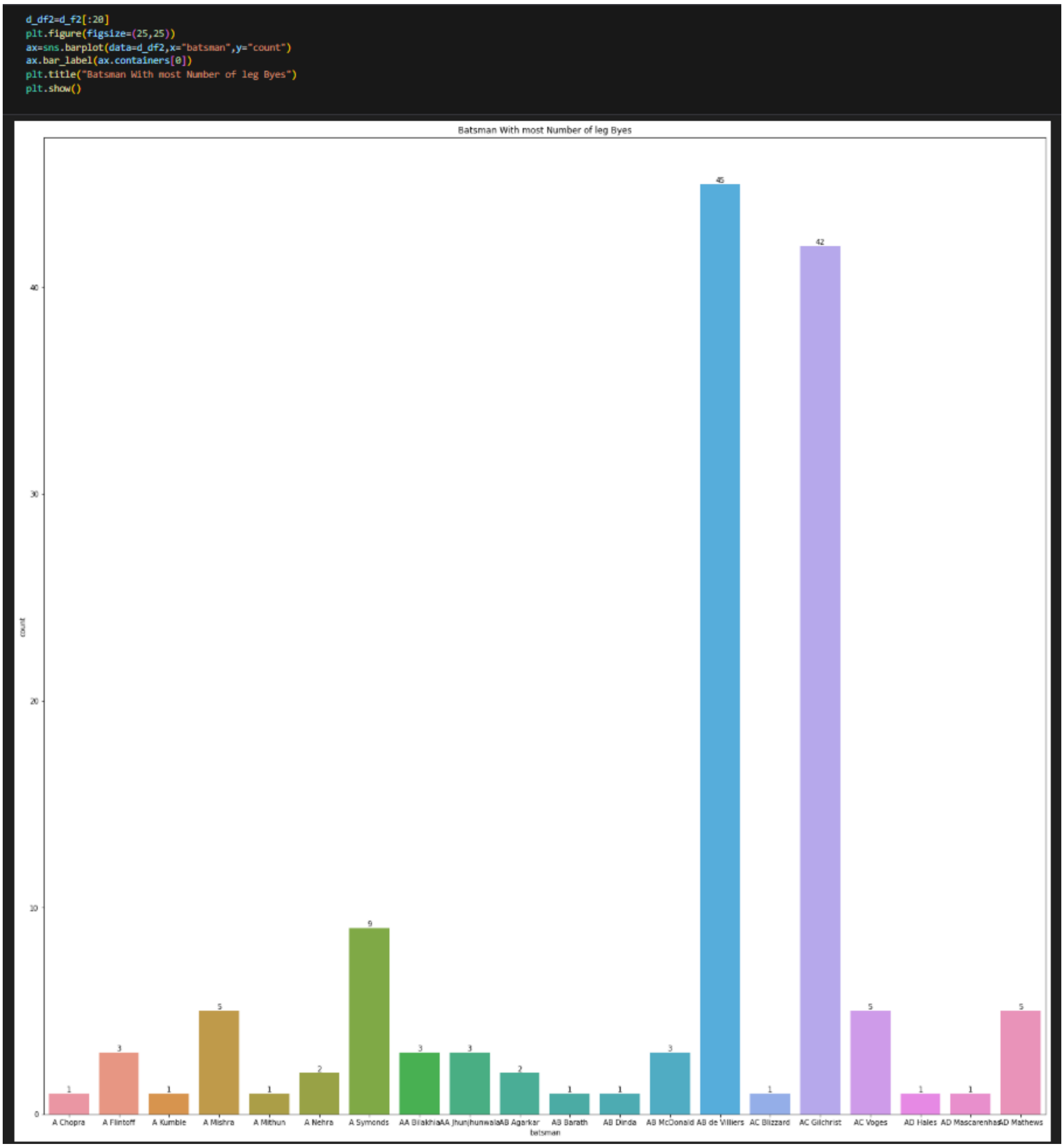
```
In [47]: plt.pie(df_ball.extras_type.value_counts().values,labels=df_ball.extras_type.v
         plt.show()
```

```
In [48]: df2=df_ball[df_ball.extras_type=="legbyes"].groupby("bowler").size().reset_ind
```

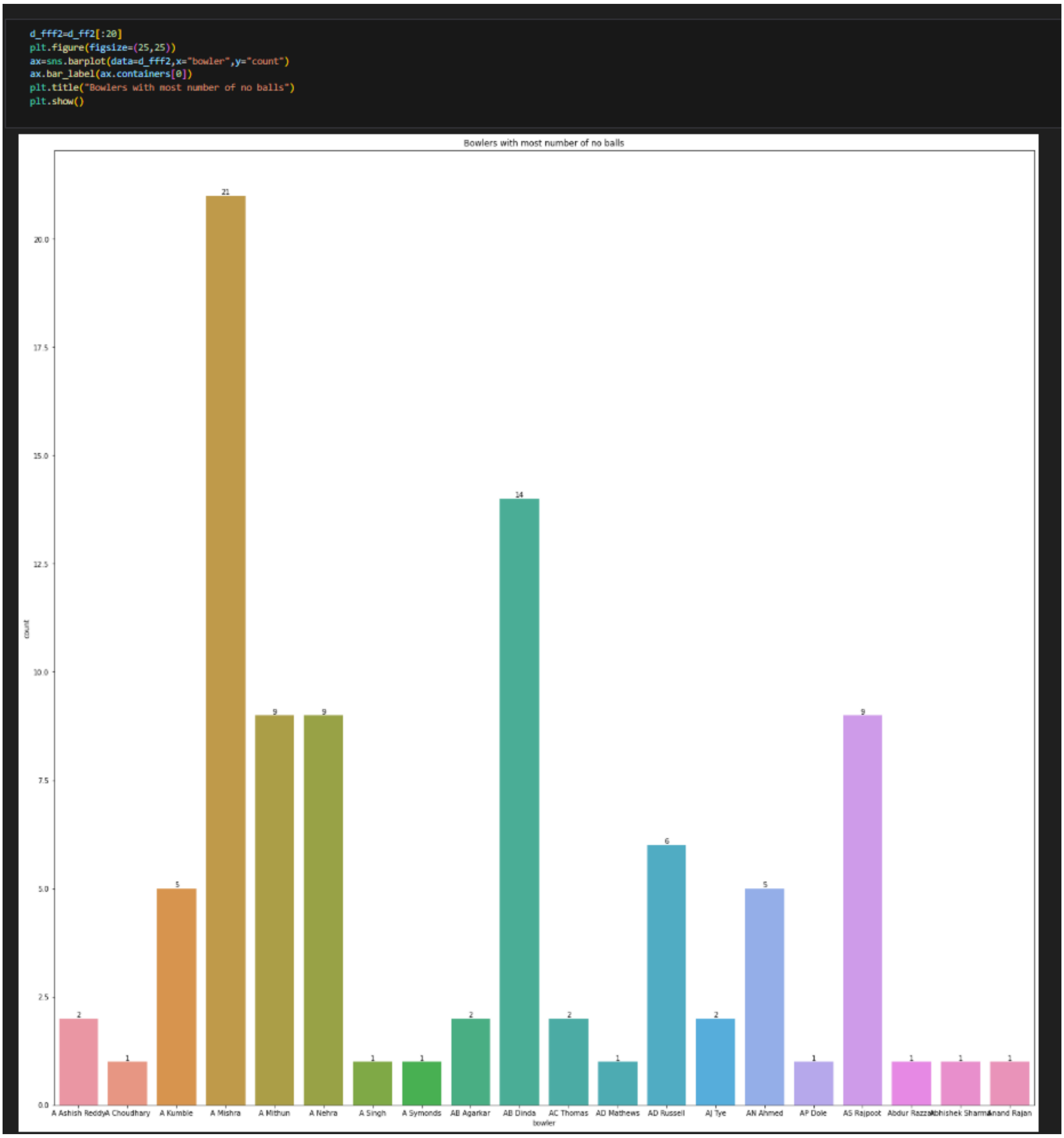
```
In [49]: plt.pie(df2["count"][:10],labels=df2["bowler"][:10],autopct="%1.2f%%")
         plt.show()
```



```
ddf2=df2[:20]
plt.figure(figsize=(20,25))
ax=sns.barplot(data=ddf2,x="bowler",y="count")
ax.bar_label(ax.containers[0])
plt.title("Bowler With most Number of leg Byes")
plt.show()
```



```
d_df2=d_f2[:20]
plt.figure(figsize=(25,25))
ax=sns.barplot(data=d_df2,x="batsman",y="count")
ax.bar_label(ax.containers[0])
plt.title("Batsman With most Number of leg Byes")
plt.show()
```



```
d_fff2=d_ff2[:20]
plt.figure(figsize=(25,25))
ax=sns.barplot(data=d_fff2,x="bowler",y="count")
ax.bar_label(ax.containers[0])
plt.title("Bowlers with most number of no balls")
plt.show()
```

Conclusion

The analysis of IPL data from 2008–2020 reveals several concrete insights about how matches are played and what factors influence outcomes. The venue analysis showed that matches are heavily concentrated in major cricketing cities such as Bengaluru, Mumbai, Delhi, and Kolkata, with a few neutral venues used only in certain seasons. Player performance trends indicate that a small group of players consistently dominate the Player of the Match awards, reflecting long-term impact and match-winning ability.

Toss analysis showed clear patterns: in many seasons, teams choosing to field first after winning the toss went on to win a large portion of matches, suggesting that conditions often favor chasing. Result margin distributions also revealed that while many matches are closely contested, certain games show large win margins, highlighting the variability in team performance.

Extras analysis produced some of the most striking findings. Bowlers like Zaheer Khan, Amit Mishra, and YS Chahal were responsible for a high number of wides and no-balls, indicating discipline issues that can significantly affect match totals. Leg-bye and bye patterns further highlighted recurring tendencies among specific bowlers and batsmen, showing where teams leak runs unintentionally.

Dismissal analysis showed clear recurring patterns in caught dismissals, including repeated fielding combinations such as CH Gayle being caught multiple times by SV Samson and Rohit Sharma being caught often by MS Dhoni. Run-out investigations revealed which bowlers or fielders were most frequently involved in dismissing top batsmen like AB de Villiers and Virat Kohli.

Overall, the results show that matches are shaped not only by big performances but also by finer details such as extras, fielding efficiency, and tactical decisions like the toss. These findings emphasize that small, repeated events at the ball-by-ball level—extras conceded, specific dismissal patterns, and bowler–batsman matchups—play a major role in determining match outcomes. The project demonstrates that detailed EDA can uncover performance weaknesses, player tendencies, and strategic patterns that are not immediately visible from match summaries alone.

GitHub link

Project link